

Компьютерная лингвистика в информационном поиске

Лекция 3

БГУ ФПМИ, 2018

План

Задачи компьютерной лингвистики

Языки

Статистические свойства языка

Кодировки

Unicode

Нормализация текста

Токенизация

Нормализация лексем

Стемминг и лемматизация

План

Задачи компьютерной лингвистики

Языки

Статистические свойства языка

Кодировки

Unicode

Нормализация текста

Токенизация

Нормализация лексем

Стемминг и лемматизация

Применение компьютерной лингвистики на различных этапах поиска

- ▶ Парсинг текстов документов.
- ▶ Сжатие индексных файлов.
- ▶ Детектирование вредоносного контента.
- ▶ Анализ текста пользовательского запроса.
- ▶ Расширение запроса.
- ▶ Извлечение признаков для ранжирования.
- ▶ Генерация сниппетов.

Обработка текста документа перед индексацией

1. Интерпретация.
2. Токенизация.
3. Нормализация

Обработка пользовательского запроса

Анализ текста запроса

- ▶ Нормализация.
- ▶ Исправление опечаток.

Расширение запроса

- ▶ Автодополнение.
- ▶ Синонимия.
- ▶ Pseudo-relevance feedback.

Зачем?

1. Отсутствие нормализации документа снижает эффективность.
 2. Поиск по непосредственному тексту запроса ограничен.
 3. Не все слова одинаково важны для поиска.
 4. Порядка 10-15% запросов сформулированы некорректно.



Задачи NLP в поиске

- ▶ Tagging (POS, NER, etc.)
- ▶ Language modeling.
- ▶ Classification (language, encoding, category, etc.)
- ▶ Clustering.
- ▶ Automatic text summarization.
- ▶ ...

План

Задачи компьютерной лингвистики

Языки

Статистические свойства языка

Кодировки

Unicode

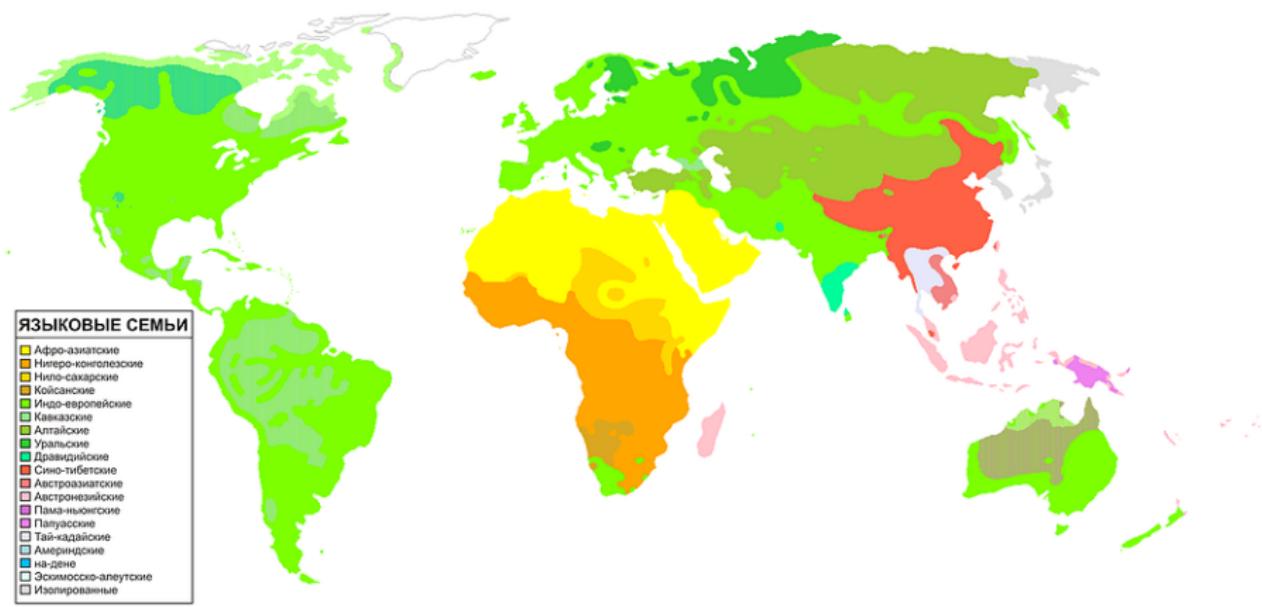
Нормализация текста

Токенизация

Нормализация лексем

Стемминг и лемматизация

Языки мира





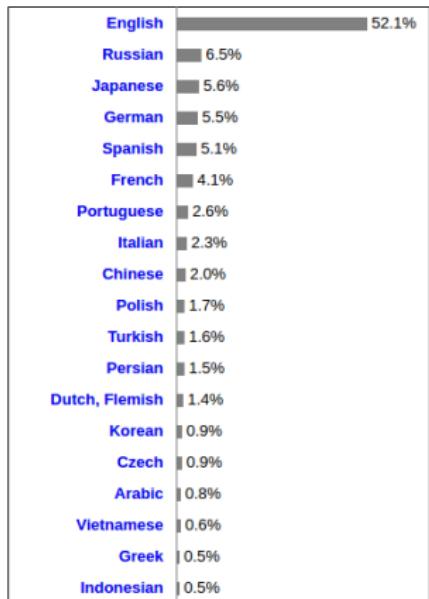
Системы письменности



Язык Интернет-пользователей

Top Ten Languages Used in the Web - June 30, 2016 (Number of Internet Users by Language)					
TOP TEN LANGUAGES IN THE INTERNET	Internet Users by Language	Internet Penetration (% Population)	Users Growth in Internet (2000 - 2016)	Internet Users % of World Total (Participation)	World Population for this Language (2016 Estimate)
English	948,608,782	67.8 %	573.9 %	26.3 %	1,400,052,373
Chinese	751,985,224	53.1 %	2,227.9 %	20.8 %	1,415,572,934
Spanish	277,125,947	61.6 %	1,424.3 %	7.7 %	450,235,963
Arabic	168,426,690	43.4 %	6,602.5 %	4.7 %	388,332,877
Portuguese	154,525,606	57.9 %	1,939.7 %	4.3 %	266,757,744
Japanese	115,111,595	91.0 %	144.5 %	3.2 %	126,464,583
Malay	109,400,982	37.8 %	1,809.3 %	3.0 %	289,702,633
Russian	103,147,691	70.5 %	3,227.3 %	2.9 %	146,358,055
French	102,171,481	25.9 %	751.5 %	2.8 %	393,892,299
German	83,825,134	88.3 %	204.6 %	2.3 %	94,973,855
TOP 10 LANGUAGES	2,814,329,132	56.6 %	848.4 %	77.9 %	4,972,343,316
Rest of the Languages	797,046,681	33.7 %	1,141.0 %	22.1 %	2,367,750,664
WORLD TOTAL	3,611,375,813	49.2 %	900.4 %	100.0 %	7,340,093,980

Язык сайт-контента



Top-10M sites, 1 March 2017

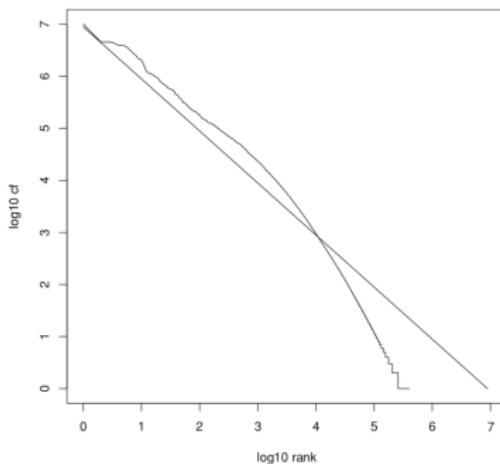
The First 100 Most Commonly Used Words 

These are the most common English words, sorted by frequency. The first 100 make up about half of all written material.

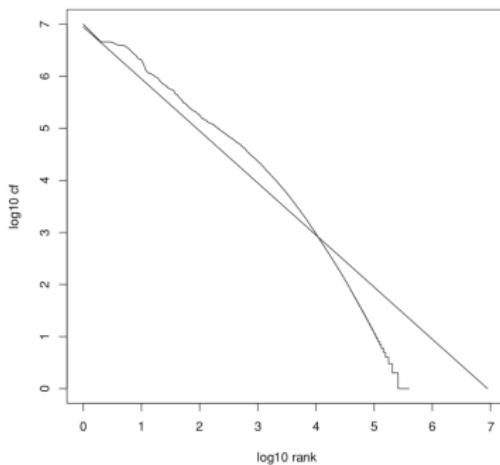
the	from	which	more
of	or	do	day
to	had	their	could
and	by	time	go
a	hot	If	come
In	word	will	did
is	but	way	number
It	what	about	sound
you	some	many	no
that	we	then	most
he	can	them	people
was	out	write	my
for	other	would	over
on	were	like	know
are	all	so	water
with	there	these	than
as	when	her	call
I	up	long	first
his	use	make	who
they	your	thing	may
be	how	see	down
at	said	him	side
one	an	two	been
have	each	has	now
this	she	look	find

Rule of 30: 30 top frequent words – 30% postings.

Закон Ципфа



Закон Ципфа



Степенной закон распределения слов:

$$cf_i \sim Ci^\alpha$$

$$\log cf_i = \log C - \alpha \log i, \quad \alpha = -1$$

Следствия из Закона Ципфа

- ▶ Ранк слова, встречающегося n_t раз:

$$r(n_t) = \frac{Cn}{n_t}$$

- ▶ $1/2$ всех слов встречается в коллекции лишь 1 раз.

Популярные слова в русскоязычных запросах

онлайн, смотреть, скачать, сайт, бесплатно, купить,
официальный, порно, фото, класс, одноклассники

Закон Хипса: оценка количества термов

Эмпирический закон Хипса

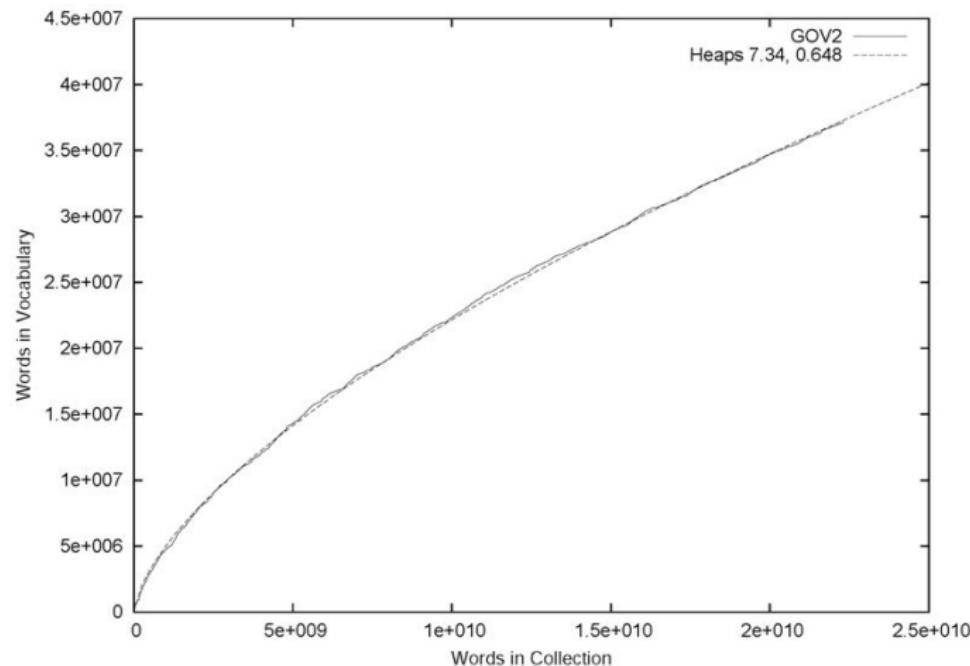
$$V(n) = kn^\beta$$

n – количество всех термов в коллекции.

$V(n)$ – количество уникальных термов в коллекции.

$10 \leq k \leq 100, 0.4 \leq \beta \leq 0.6$ – параметры.

Закон Хипса



План

Задачи компьютерной лингвистики

Языки

Статистические свойства языка

Кодировки

Unicode

Нормализация текста

Токенизация

Нормализация лексем

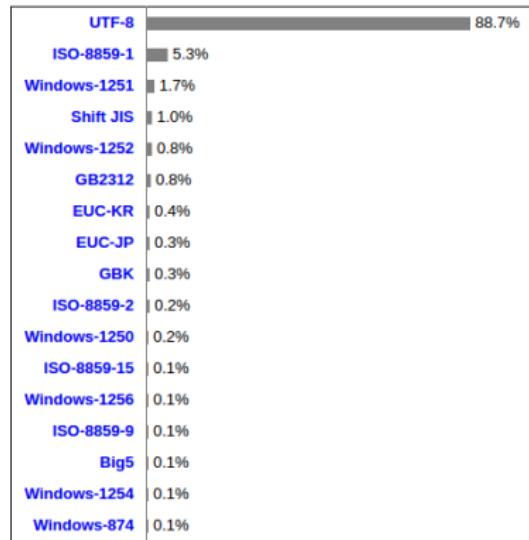
Стемминг и лемматизация

Кодировка

Таблица, сопоставляющая каждому символу алфавита последовательность из символов другого алфавита.

Например, азбука Морзе, сигнальные флаги. В компьютере это последовательности бит.

Доли кодировок



Top-10M sites, 1 March 2017

Unicode

- ▶ Стандарт, позволяющий представить знаки практически всех письменных языков.
- ▶ Каждому символу соответствует уникальный код $U+nnnn$.
- ▶ Машинное представление кодов определяется кодировками из семейства UTF (UTF-8, UTF-16BE, UTF-16LE, UTF-32BE, UTF-32LE).

UTF-8

Обратная совместимость с ASCII.

Отображение кодов Unicode в байты UTF-8:

0x00000000 — 0x0000007F: 0xxxxxxxx

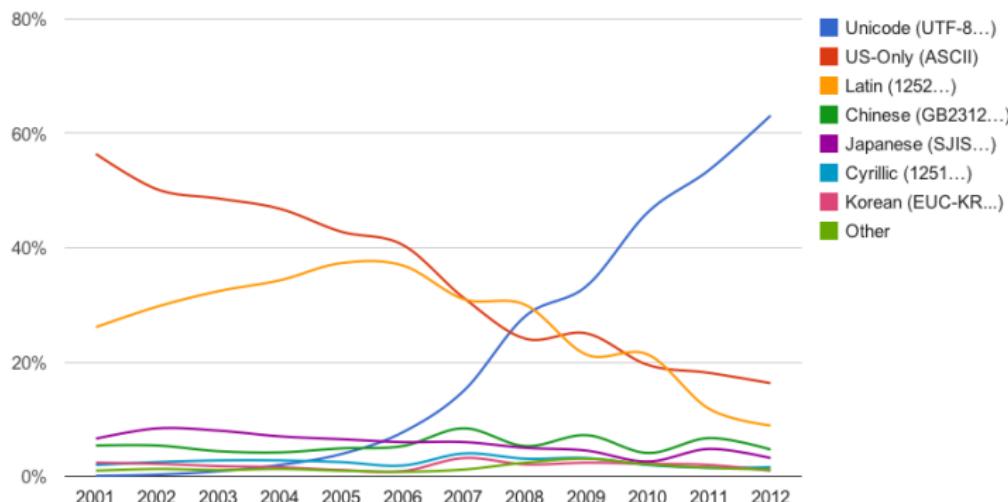
0x00000080 — 0x000007FF: 110xxxxx 10yyyyyy

0x00000800 — 0x0000FFFF: 1110xxxx 10yyyyyy 10zzzzzz

0x00010000 — 0x001FFFFF: 11110xxx 10yyyyyy 10zzzzzz

10tttttt

Рост Unicode



Как распознать язык/кодировку?

- ▶ Эмпирические правила (наличие характерных служебных слов).
- ▶ Информационные методы (сжимаемость данных).
- ▶ Статистические методы:
 - ▶ классификатор поверх n -грамм;
 - ▶ соответствие n -граммным моделям, обученным для различных языков;
 - ▶ ...

План

Задачи компьютерной лингвистики

Языки

Статистические свойства языка

Кодировки

Unicode

Нормализация текста

Токенизация

Нормализация лексем

Стемминг и лемматизация

Первичная обработка текста

- ▶ Приведение к нижнему регистру.
- ▶ Удаление знаков препинания.
- ▶ Удаление стоп-слов.

Особенности токенизации

- ▶ Апостроф внутри лексемы ('o'clock', 'O'Sullivan').
- ▶ Разные варианты написания сокращений ('IBM' vs 'I.B.M.').
- ▶ Фразы ('Los Angeles').
- ▶ Обработка дефиса ('lower-case' vs 'lowercase').

Как решать?

- ▶ Частотный словарь n -грамм.
- ▶ Синтаксический анализатор.
- ▶ Машинное обучение с размеченным корпусом.
- ▶ Теггеры (POS, NER, etc.).

Нормализация лексем

приведение лексем к канонической форме путем задания классов эквивалентности.

Слишком сильная нормализация ведет к уменьшению точности при росте полноты.

Проблема пересекающихся классов

Пример: акронимы, по написанию совпадающие с другими словами ('US', 'U.S.' vs 'us').

Решение: двойная индексация ('United States' и 'we').

Способы задания эквивалентных классов

1. Словарная морфология.
2. Стемминг.
3. Кластеризация.
4. Tagging.
5. n -граммное индексирование.

Нормализация путем выделения основ

Цель: привести словоформы и производные формы слова к общей основе.

Нормализация путем выделения основ

Цель: привести словоформы и производные формы слова к общей основе.

Лемматизация

Точный морфологический анализ с использованием лексикона, в результате которого удаляются только флексивные морфемы и возвращается словарная форма слова (лемма).

Нормализация путем выделения основ

Цель: привести словоформы и производные формы слова к общей основе.

Лемматизация

Точный морфологический анализ с использованием лексикона, в результате которого удаляются только флексивные морфемы и возвращается словарная форма слова (лемма).

Стемминг

Эвристический процесс, в ходе которого от слов отбрасываются окончания согласно набору правил.

Алгоритмы стемминга

Пример. Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Алгоритм Ловинса. such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Алгоритм Портера. such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Алгоритм Пейса-Хаска. such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

Стеммер Портера

- ▶ 5 последовательных этапов применения правил.
- ▶ На каждом этапе применяются соглашения по выбору правил (например, выбор по применимости к самому длинному суффиксу).

Правила первого этапа:

SSES → SS, (caresses → caress)

IES → I, (ponies → poni)

SS → SS, (caress → caress)

S → , (cats → cat)

Стемминг

- ▶ Зачастую выгоднее точной лемматизации.
- ▶ Существенный выигрыш в эффективности дает на финском, испанском.

Snowball – фреймворк стемминга для различных языков.